В. Г. В а с и л ь е в (Москва, ИПИ РАН)**. Автоматическое выделение значимых фрагментов в текстах.**

Доклад посвящен рассмотрению подходов к выделению значимых фрагментов в текстах при решении задач автоматической классификации. С точки зрения используемых моделей текстовых данных он развивает идеи, которые ранее представлялись автором на сессиях 6-го и 7-го симпозиумов по прикладной и промышленной математике ([1], [2]).

Формально задача определяется следующим образом. Пусть $\omega_1, \ldots, \omega_k$ — рубрики иерархического классификатора, задающие темы, которые представляют интерес и которые требуется автоматически выделять в текстах, $X = (X_1, \ldots, X_n)$ — текст на естественном языке, состоящий из n предложений, X_i — вектор весов информационных признаков в предложении $i = 1, \ldots, n$. Требуется для каждой рубрики ω_j, $j = 1, \ldots, k$, определить факт наличия в тексте информации по ней и в случае положительного решения найти предложения ей соответствующие.

Основной сложностью при выделении фрагментов в текстах является то, что в общем случае оценку принадлежности отдельных предложений к рубрике можно проводить только с учетом контекста их употребления в тексте. Это связано с тем, что лексического состава одного предложения может быть не достаточно для принятия решения о его принадлежности к рубрике, а лексический состав большого фрагмента, включающего данное предложение, уже может быть избыточным.

В докладе рассматриваются два новых подхода к выделению фрагментов. В первом подходе осуществляется классификация фрагментов в рамках метода «скользящего окна», а для уточнения границ фрагментов используется скрытая марковская модель. Во втором подходе вводится новая мера близости предложений к рубрикам, основанная на использовании результатов классификации специально построенного покрытия предложений фрагментами текста. При этом покрытие строится таким образом, чтобы алгоритм имел бы линейную сложность от числа предложений в тексте и в то же время обеспечивал точное определение границ фрагментов.

Общим элементом приведенных подходов является использование иерархического комбинированного классификатора, который основан на совместном использовании вероятностных (байесовских) методов классификации, методов классификации на основе расстояний и методов классификации на основе правил, задаваемых экспертами. Формулируются основные проблемы, которые возникают при обработке реальных данных на этапах обучения и классификации. Приводится описание технологии классификации текстов, ориентированной на решение задачи выделения фрагментов с учетом наличия различных аномалий в исходных данных.

В заключительной части даются примеры обработки массивов новостных сообщений и результаты сравнительного анализа качества автоматической классификации текстов с учетом и без учета выделения фрагментов. Оценка качества осуществляется путем сравнения с эталонной классификацией. Делаются выводы и определяются перспективные направления дальнейших исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Васильев В. Г.*, *Ефременкова М. В.*, *Кривенко М. П.* Библиотека процедур классификации текстовых данных. — Обозрение прикл. и промышл. матем., 2006, т. 13, в. 1, с. 86–87.
2. *Кривенко М. П.*, *Васильев В. Г.* Модели кластерной структуры массивов текстовых данных. — Обозрение прикл. и промышл. матем., 2005, т. 12, в. 3, с. 743.